Erschienen in: Kallmeyer, Werner/Zifonun, Gisela (Hrsg.): Sprachkorpora – Datenmengen und Erkenntnisfortschritt. - Berlin/New York: de Gruyter, 2007. S. VII-X. (Jahrbuch des Instituts für Deutsche Sprache 2006)

Vorwort

Das Beschaffen, Dokumentieren und Archivieren von Sprachdaten gehört seit jeher zu den arbeitsaufwändigen, aber unverzichtbaren Tätigkeiten sprachwissenschaftlicher Forschung. Aber seit einer Reihe von Jahren widmen sich weltweit bedeutende wissenschaftliche Unternehmungen dem Aufbau großer Korpora. Die leitenden Gesichtspunkte sind dabei unterschiedlich, sie zielen u. a. auf die Sammlung und Dokumentation von schwer zu beschaffenden Sprachmaterialien, z. B. von bedrohten Sprachen (vgl. das entsprechende Programm der VW-Stiftung), auf die Weiterentwicklung der Theorie von Daten und Korpora ebenso wie auf die Entwicklung von Verfahren der Korpusanalyse (vgl. u. a. den SFB 441 „Linguistische Datenstrukturen" in Tübingen und das Kooperationsprojekt „Nachhaltigkeit linguistischer Daten" der SFBs 441, 538 und 632). So kann es auch nicht erstaunen, dass der linguistische Tagungskalender dieser Jahre mehrere Tagungen mit einer korpusorientierten Ausrichtung aufweist (so die Jahrestagung 2006 der Deutschen Gesellschaft für Sprachwissenschaft „Sprachdokumentation und Sprachbeschreibung", die Jahrestagung 2006 des IDS „Sprachkorpora – Datenmengen und Erkenntnisfortschritt" und der Kongress 2007 der International Pragmatics Association „Language data, corpora, and computational pragmatics").

In dieser Intensivierung von Korpusaufbau und Korpusanalyse wird die technische Entwicklung sichtbar. Die Sprachwissenschaften verfügen heute über korpustechnologische Werkzeuge, die einen relativ komfortablen Umgang mit großen Datenmengen erlauben. Mit der Verfügbarkeit von geschriebenen Texten in elektronischer Form wird es leichter, große Textkorpora zusammenzustellen – wenn man einmal von den teilweise dornigen Fragen des Copyrights absieht. Die Größe von schriftlichen Korpora, soweit sie sich auf öffentliche Texte konzentrieren, wächst geradezu explosionsartig im Vergleich mit den Korpora der 1970er Jahre. Für die gesprochene Sprache ist die Lage in dieser Hinsicht allerdings völlig anders. Die Aufnahme- und Speichertechnik für Ton und auch Video ermöglicht heute ebenfalls auf einfache Weise Sprachaufnahmen „im Feld", d. h. in natürlichen Kommunikationssituationen. Außerdem wird öffentlich gesprochene Sprache von den Medien Rundfunk und Fernsehen in großer Menge verbreitet, in etwa vergleichbar dem öffentlichen Markt der gedruckten Texte. Hinzu kommt, dass die verfügbaren und finanzierbaren Speichermedien für Ton und Bild heute eine „massenhafte" Speicherung solcher Daten erlauben. Aber die Korpora sind kaum

linguistisch auswertbar ohne Verschriftlichung des Gesprochenen. Die Transkription ist gleichsam das Nadelöhr, durch das mündliche Sprachdaten gezwängt werden müssen, um als Analysegrundlage dienen zu können. Angesichts des Arbeitsaufwandes, den eine brauchbare Verschriftlichung gesprochener Sprache erfordert, wundert es nicht, dass die mündlichen Korpora ungleich langsamer wachsen als die schriftlichen.

In der verstärkten Zuwendung zu korpuslinguistisch gestützten Untersuchungen zeigt sich teilweise auch eine Veränderung der theoretischen und methodischen Orientierung der Linguistik. Die Situation stellt sich in dieser Hinsicht allerdings für die verschiedenen linguistischen Teilbereiche unterschiedlich dar.

Die Beiträge des vorliegenden Jahrbuchs konzentrieren sich insbesondere auf theoretische und methodische Fragen zum Aufbau und zur Nutzung großer Sprachkorpora. Dabei treten sowohl grundlegende Fragen der linguistischen Arbeit mit Korpora hervor als auch bereichsspezifische Probleme.

Aus einer Perspektive der Grundlagenreflexion expliziert **Christian Lehmann** die Bedingungen für Linguistik als empirische Wissenschaft durch die Bestimmung von Daten, die Anforderungen an die Dokumentation und den produktiven Umgang mit Korpora. Die Auseinandersetzung mit der Kernfrage, ob, in welchem Sinne und wie die Linguistik eine empirische Wissenschaft ist, führt **Anke Lüdeling** fort mit der auf die Korpuslinguistik bezogenen Argumentation, wonach die Grundlage jeder quantitativen Analyse die qualitative Analyse oder Kategorisierung der Daten ist.

Für die grammatische Forschung erweisen sich zunehmend die in großen Korpora anzutreffenden Variationsphänomene und – zumindest auf den ersten Blick – auffälligen und irregulär erscheinenden Äußerungsstrukturen als wichtige Informationsquelle. Insofern hat sich eine Entwicklung hin zur korpusbasierten Arbeit etabliert. Die Erhebung von Grammatikalitätsurteilen ist ein wichtiges Instrument der grammatischen Forschung, wie **Sam Featherston** anhand der experimentellen Erhebung introspektiver Urteile in numerischer Form auf einer Intervallskala und ihrer Quantifizierung demonstriert. Auf der anderen Seite zeigt **Stefan Müller** am Beispiel von drei syntaktischen Phänomenen (Partikelverben, Extraposition und Subjazenz sowie mehrfache Vorfeldbesetzung) den Wert von Korpusabfragen, weil den Sprechern ihr sprachliches Wissen nicht immer zugänglich ist. Ein dritter Aspekt der neueren grammatischen Forschung ist die Arbeit mit syntaktisch annotierten Korpora (vgl. auch die Präsentation des TIGER-Korpus während der Tagung). Die Vorteile und Schwierigkeiten dieses Vorgehens für die historische Sprachforschung zeigt **Ulrike Demske** anhand des Aufbaus einer Baumbank für das Frühneuhochdeutsche.

Für die Lexikographie gehört es heute zum Standard, mit Unterstützung durch große maschinenlesbare Korpora zu arbeiten. Vor diesem Hintergrund diskutiert **Annette Klosa** anhand wichtiger Vorhaben korpusgestützter Lexiko-

graphie die Auswirkungen auf die Strukturierung der Wörterbuchartikel und die Wahl der Beispiele. **Jörg Asmussen** zeigt den Wert von korpuslinguistischen Verfahren für die Analyse von Bedeutungsparaphrasen in Wörterbüchern und ihre Optimierung am Beispiel eines korpusgestützten Wörterbuchs der dänischen Sprache.

Die Beiträge zur Untersuchung von Daten gesprochener Sprache gehen von einer ethnolinguistischen, soziolinguistischen beziehungsweise pragmatischen Perspektive aus. Bei der Arbeit mit kulturell fremden Daten tritt u. a. das Problem einer angemessenen Dokumentation des für die Analyse notwendigen kulturellen Hintergrundwissens hervor. Diese zusätzliche Anforderung gegenüber der Arbeit mit Daten aus vertrauten kulturellen Kontexten führt **Gunter Senft** an Beispielmaterial von Trobriand-Insulanern aus Papua-Neuguinea vor. Grundsätzlich, wenn auch nicht so zugespitzt, besteht dieser Bedarf an reichen Metadaten auch bei Feldforschungsdaten aus dem deutschen Sprachraum, z. B. bei regionalsprachlichen Untersuchungen, worauf auch **Alexandra Lenz** hinweist. Sie gibt einen Überblick über die verfügbaren regionalsprachlichen Korpora des Deutschen und demonstriert Möglichkeiten und gegenwärtig noch bestehende Beschränkungen für die variationslinguistische Auswertung. **Werner Kallmeyer** zeigt auf, welche besonderen Anforderungen sich bei Korpora sprachlicher Interaktion für die Entwicklung von Verfahren der computergestützten Volltextrecherche auf COSMAS-Basis stellen.

Die beiden folgenden Beiträge diskutieren Fortschritte und Aussichten der Korpusanalyse aus informatischer Sicht. Die Diskussion auf diesem Gebiet spiegelt noch einmal die in den linguistischen Beiträgen bereits aufscheinende Frage, ob und wann die Arbeit mit „reinen Daten" oder die Anreicherung von Korpora mit Interpretationen (Annotation) sinnvoller ist und welche Möglichkeiten quantitative und wissensbasierte Auswertungsverfahren bieten. **Thorsten Brants** präsentiert Verfahren der statistisch basierten maschinellen Übersetzung, eine für Sprachwissenschaftler eher ungewohnte, aber anregende Perspektive. **Michael Strube**, **Margot Mieskes** und **Christoph Müller** schließlich locken mit der Erfüllung einer Wunschvorstellung von Protokollschreibern wie Gesprächsforschern, die diese aber immer wieder verdrängen mussten: „Gesprächsprotokolle auf Knopfdruck".

Das Vortragsprogramm der Jahrestagung 2006 wurde durch einen Nachmittag mit korpustechnologischen Präsentationen ergänzt; Informationen dazu finden sich auf den Internetseiten des IDS unter der Adresse <www.ids-mannheim.de/org/tagungen/jt2006/praesentation.html>.

Den Abschluss der Tagung bildete eine Podiumsdiskussion zum Thema „Varianten im Korpus". Unter der Moderation von Norbert R. Wolf diskutierten Ulrike Hass, Christian Lehmann, Anke Lüdeling, Christian Mair und Gisela Zifonun miteinander und dem Publikum darüber, in welchem Umfang Variation in einem als standardsprachlich oder standardsprachennah konzipierten großen Korpus etwa des Deutschen, das für unterschiedliche Frage-

stellungen genutzt werden soll, zu berücksichtigen ist und in welcher Weise Architektur und Design des Korpus dem Faktor Varianz Rechnung tragen können. Ein solches Korpus kann, darin stimmten die Diskutanten überein, nicht repräsentativ für den Sprachgebrauch einer Sprachgemeinschaft sein. Wohl aber kann es als ausgewogenes oder exemplarisches Korpus eine begründete Auswahl aus der Vielfalt kommunikativer Gattungen vertreten, die in angemessenem Verhältnis zueinander berücksichtigt werden. In einem solchen Korpus wird Varianz, „nach der man nicht gefragt hat", in jedem Fall auf der Ebene der Rohdaten vorhanden sein. Varianz – etwa in der phonetischen Realisierung – wird auf abstrakteren Ebenen der Datenrepräsentation gegebenenfalls reduziert werden (müssen). Bei Variation auf der Ebene der Morphosyntax oder der Satzverknüpfung stellt sich häufig die Frage, ob eine unter normativer Sehweise „anstößige" Variante als Performanzfehler oder als zu Unrecht stigmatisierte Ausdrucksform oder gar als Indiz für Sprachwandelprozesse zu betrachten ist.

Werner Kallmeyer
Gisela Zifonun